

**REMARKS**

Claims 8, 9 and 12 have been canceled as depending from a non-existing claim, claim 7. Claim 2 has been amended to delete sequences related to non-elected inventions. Claim 11 has been amended to correct dependence from canceled claim 9 to claim 10 and provide proper antecedent for the term “composition” in the claim. No new matter is added by any of these amendments, and entry of the amendments is requested..

The Examiner stated that applicants arguments regarding the traversal of the Restriction Requirement limiting the examination of claims to the single elected sequence of SEQ ID NO:6 are non-persuasive, and the requirement is deemed proper and made FINAL.

The Examiner stated, however, that the polypeptide encoded by SEQ ID NO:6, SEQ ID NO:22, has been searched, and that invention 59, claims 15 and 16 drawn to SEQ ID NO:22, has been rejoined with invention 6, drawn to SEQ ID NO:6. Claims 1-3 and 12-16 as drawn to SEQ ID NO:6 or 22 are under examination. Applicants reserve the right to prosecute non-elected subject matter in subsequent divisional applications.

**35 U.S.C. § 101, Rejection of Claims 1-3 and 12-16**

The Examiner has rejected claims 1-3 and 12-16 under 35 U.S.C. § 101, because the claimed invention lacks patentable utility. The Examiner stated that the specification teaches that the polynucleotide of SEQ ID NO:6, which encodes the polypeptide of SEQ ID NO:22, is a matrix remodeling gene because it is co-expressed with known matrix-remodeling genes. However, the Examiner stated, coexpression of genes does not provide evidence regarding the function of the encoded gene product. Further, even if the gene encodes a protein involved in matrix-remodeling, its role or activity in matrix-remodeling has not been disclosed. The assertion that SEQ ID NO:6 and its encoded protein SEQ ID NO:22 are involved in matrix remodeling because the gene is coexpressed with known matrix remodeling genes lacks basis for utility because coexpression of a gene does not correspond to gene or gene product function.

The Examiner stated further that the polynucleotide sequence consisting of SEQ ID NO:6 may have utility because it encodes a protein having utility. However, applicants assertions that the protein of SEQ ID NO:22 resembles RH1 and RH2 opsins (see page 29, paragraph 1) is not support by a

review of the art demonstrating that opsins are G-protein coupled receptors comprising approximately 350-400 amino acids. See, for example, Cowman et al., Kaushal et al., Pasqualetti et al., and Zuker et al., cited at page 4 of the Office Action.

The asserted utilities are general utilities and do not form a substantial utility because further research is needed to identify or reasonably confirm a real world context of use for SEQ ID NO6 and its encoded protein SEQ ID NO:22.

Applicants Response

Applicants disagree that the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility. The Examiner's rejection is primarily based on the mistaken premise that applicant's assertion of utility for the claimed polynucleotides and proteins requires their having a specific role or activity in matrix-remodeling.

As the title of the invention clearly states, the claimed invention is directed to "Polynucleotides Coexpressed with Matrix-Remodeling Genes" and their encoded proteins. The invention employs a method for identifying biomolecules that are associated with a specific disease, regulatory pathway, subcellular compartment, cell type, etc., known as "guilt by association", and uses known marker genes for a condition, disease or disorder to identify surrogate markers, polynucleotides or proteins that are coexpressed in the same condition, disease or disorder. See specification, at page 6, second paragraph. In the instant case, the method was employed to identify SEQ ID NOs:1-20, and their encoded polypeptides, SEQ ID NOs:21-23, that are highly significantly coexpressed with at least two of twenty-one known genes and their gene products that are involved in matrix-remodeling and associated with diseases involving matrix-remodeling. See, table at pages 23-25 of the specification, that describes the functions and disease associations of these twenty-one known genes. The data in Table 4, page 27 of the specification demonstrates that the "strong association" of these twenty novel SEQ ID NOs: was "distilled" from an analysis of some 41,000 genes that identified them as having an extremely low "due-to-chance" probability of less than  $10^{-7}$  for their association with the known matrix-remodeling genes. See specification, at page 8, second paragraph, and references therein.

Thus, while there is a substantial likelihood that these 20 novel genes that are coexpressed with known matrix-remodeling genes are, themselves, involved in matrix re-modeling (see, in particular, Walker and Volkmuth (1999) Prediction of gene function by genome-scale expression analysis:

prostate-associated genes. Genome Res 9:1198-1203, cited at page 6 of the specification), their use as surrogate markers for the known genes in the diagnosis or evaluation of therapies for diseases associated with matrix-remodeling does not require that they function in any particular aspect of matrix-remodeling. Thus, the Examiner's allegation that the specific role or activity for the claimed polynucleotide, or its encoded protein, in matrix-remodeling be disclosed in order for it to be used for its asserted purpose is unfounded.

Applicants further submit that the specification and art of record suggests a well established utility for the claimed polynucleotides and proteins. The specification and art of record discloses that the claimed nucleic acids may be used as probes in a variety of gene and protein expression monitoring applications, and that such gene expression monitoring applications are highly useful in drug development and in toxicology testing that was well known at the time the application was filed. See specification, at page 14, line 17 through page 16, line 21.

In support of such a well-established utility, applicants hereby submit three expert Declarations under 37 C.F.R. § 1.132, with respective attachments, and ten (10) scientific references filed before the October 9, 1998 priority date of the instant application. The Rockett Declaration, Iyer Declaration, Bedilion Declaration, and the ten (10) references fully establish that, prior to the October 9, 1998 priority date of the instant application, it was well-established in the art that:

polynucleotides derived from nucleic acids expressed in one or more tissues and/or cell types can be used as hybridization probes -- that is, as tools -- to survey for and to measure the presence, the absence, and the amount of expression of their cognate gene;

with sufficient length, at sufficient hybridization stringency, and with sufficient wash stringency -- conditions that can be routinely established -- expressed polynucleotides, used as probes, generate a signal that is specific to the cognate gene, that is, produce a gene-specific expression signal;

expression analysis is useful, *inter alia*, in drug discovery and lead optimization efforts, in toxicology, particularly toxicology studies conducted early in drug development efforts, and in phenotypic characterization and categorization of cell types, including neoplastic cell types;

each additional gene-specific probe used as a tool in expression analysis provides an additional gene-specific signal that could not otherwise have been detected, giving a more comprehensive, robust, higher resolution, statistically

more significant, and thus more useful expression pattern in such analyses than would otherwise have been possible;

biologists, such as toxicologists, recognize the increased utility of more comprehensive, robust, higher resolution, statistically more significant results, and thus want each newly identified expressed gene to be included in such an analysis;

nucleic acid microarrays increase the parallelism of expression measurements, providing expression data analogous to that provided by older, lower throughput techniques, but at substantially increased throughput;

accordingly, when expression profiling is performed using microarrays, each additional gene-specific probe that is included as a signaling component on this analytical device increases the detection range, and thus versatility, of this research tool;

biologists, such as toxicologists, recognize the increased utility of such improved tools, and thus want a gene-specific probe to each newly identified expressed gene to be included in such an analytical device;

the industrial suppliers of microarrays recognize the increased utility of such improved tools to their customers, and thus strive to improve salability of their microarrays by adding each newly identified expressed gene to the microarrays they sell;

it is not necessary that the biological function of a gene be known for measurement of its expression to be useful in drug discovery and lead optimization analyses, toxicology, or molecular phenotyping experiments;

failure of a probe to detect changes in expression of its cognate gene does not diminish the usefulness of the probe as a research tool; and

failure of a probe completely to detect its cognate transcript in any single expression analysis experiment does not deprive the probe of usefulness to the community of users who would use it as a research tool.

The Patent Examiner does not dispute that the claimed polynucleotide can be used as a probe in cDNA microarrays and used in gene expression monitoring applications. Instead, the Patent Examiner contends that the claimed combination of polynucleotides cannot be useful without precise knowledge of their biological activities. See Office Action at page 3. Applicants submit that such a

position is not supported by the law as it applies to the utility requirement under 35 U.S.C. §§ 101 and 112.

**I. The applicable legal standard**

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is “practically useful,” *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a “specific benefit” on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

An invention is “useful” under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) (“to violate Section 101 the claimed device must be totally incapable of achieving a useful result”); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention “is incapable of serving any beneficial end”).

*Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: “[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility.” *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See *Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a “nebulous expression” such as “biological activity” or “biological properties” that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be “substantial.” *Brenner*, 383 U.S. at 534. A “substantial” utility is a practical, “real-world” utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a “well-established” utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examining Procedure at § 706.03(a). Only if there is no “well-established” utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a “substantial likelihood” of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

**II. Use of the claimed polynucleotides in disease detection and diagnosis and in toxicology testing are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph**

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are “well-established” uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application’s specification. These uses are explained, in detail, in the Rockett Declaration, Iyer Declaration, and Bedilion Declaration accompanying this brief. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

**A. The use of the claimed polynucleotides for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer “specific benefits” to the public**

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying Rockett Declaration, Iyer Declaration, and Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotide.

In his Declaration, Dr. Rockett explains the many reasons why a person skilled in the art in 1998 would have understood that any expressed polynucleotide is useful for a number of gene expression monitoring applications, *e.g.*, in cDNA microarrays, in connection with the development of drugs and the monitoring of the activity of such drugs. (Rockett Declaration at, *e.g.*, ¶¶ 10-18).

It is my opinion, therefore, based on the state of the art in toxicology at least since the mid-1990s . . . that disclosure of the sequence of a new gene or protein, with or without knowledge of its biological function, would have been sufficient information for a toxicologist to use the gene and/or protein in expression profiling studies in toxicology.<sup>1</sup> [Rockett Declaration, ¶ 18.]

In his Declaration, Dr. Bedilion explains why a person of skill in the art in 1998 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays. (Bedilion Declaration, *e.g.*, ¶¶ 4-7.) In his Declaration, Dr. Iyer explains why a person of skill in the art in 1998 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays, stating that “[t]o provide maximum versatility as a research tool, the microarray should include – and as a biologist I would want my microarray to include – each newly identified gene as a probe.” (Iyer Declaration, ¶ 9.)

---

<sup>1</sup> “Use of the words ‘it is my opinion’ to preface what someone of ordinary skill in the art would have known does not transform the factual statements contained in the declaration into opinion testimony.” *In re Alton*, 37 USPQ2d 1578, 1583 (Fed. Cir. 1996).

In addition, Dr. Rockett explains in his Declaration that “there are a number of other differential expression analysis technologies that precede the development of microarrays, some by decades, and that have been applied to drug metabolism and toxicology research, including: (1) differential screening; (2) subtractive hybridization, including variants such as chemical cross-linking subtraction, suppression-PCR subtractive hybridization and representational difference analysis; (3) differential display; (4) restriction endonuclease facilitated analyses, including serial analysis of gene expression (SAGE) and gene expression fingerprinting and (5) EST analysis.” (Rockett Declaration, ¶ 7.)

Nowhere does the Patent Examiner address the fact that, as described on page 15 of the Walker application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays – probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon v. Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 (“Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)” (emphasis added)).

Literature reviews published shortly before or after the filing of the Walker application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis,



recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

\* \* \*

Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

\* \* \*

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original)

Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, Xenobiotica 29:655-691 (July 1999) (Rockett Declaration, Exhibit C).

In a pre-October 1998 article, Lashkari et al. state explicitly that sequences that are merely “predicted” to be expressed (predicted Open Reading Frames, or ORFs) – the claimed invention in fact is known to be expressed – have numerous uses:

Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons– they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay. (emphasis added)

Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (Aug. 1997) (Rockett Declaration, Exhibit F).

**B. The use of polynucleotides coding for polypeptides expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now “well-established”**

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, e.g., as described by Bedilion, Rockett, and Iyer in their Declarations.

Toxicology testing is now standard practice in the pharmaceutical industry. See, e.g., John C. Rockett et al., *supra*:

Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Rockett Declaration, Exhibit C, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999) (Reference No.1); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000) (Reference No. 2).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, Environ. Health Perspec.107:681-685 (1999) (Reference No. 3). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 4), indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

Further evidence of the well-established utility of all expressed polypeptides and polynucleotides in toxicology testing is found in U.S. Pat. No. 5,569,588 (Reference No. 9e) and published PCT applications WO 95/21944 (Reference No. 9a), WO 95/20681 (Reference No. 9b), and WO 97/13877 (Reference No. 9g).

WO 95/21944 ("Differentially expressed genes in healthy and diseased subjects"), published August 17, 1995, describes the use of microarrays in expression profiling analyses, emphasizing that *patterns* of expression can be used to distinguish healthy tissues from diseased tissues and that *patterns* of expression can additionally be used in drug development and toxicology studies, without knowledge of the biological function of the encoded gene product. In particular, and with emphasis added:

The present invention involves . . . methods for diagnosing diseases . . . characterized by the presence of [differentially expressed] . . . genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/ polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST. . . . Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of diseases based on differential expression of genes of unknown function. . . . [abstract]

The method [of the present invention] involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences . . . and a defined set of oligonucleotide/polynucleotide[] . . .

immobilized on a support. Those defined [immobilized] oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined by the collection of partial cDNA sequences (ESTs). [page 2]

The present invention meets the unfilled needs in the art by providing methods for the . . . use of gene fragments and genes, even those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. [page 4]

Yet another aspect of the invention is that it provides . . . a means for . . . monitoring the efficacy of disease treatment regimes including . . . toxicological effects thereof." [page 4]

It has been appreciated that one or more differentially identified EST or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the EST[] identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. [page 4]

As used herein, the term 'disease' or 'disease state' refers to any condition which deviates from a normal or standardized healthy state in an organism of the same species in terms of differential expression of the organism's genes. . . [whether] of genetic or environmental origin, for example, an inherited disorder such as certain breast cancers. . . .[or] administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects gene expression. [page 5]

As used herein, the term 'solid support' refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method . . . [and includes, inter alia,] nitrocellulose, . . . glass, silica. . . . [page 6]

By 'EST' or 'Expressed Sequence Tag' is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides. . . . [page 6]

One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000 - 100,000 in an animal such as a human. [page 7]

The lengths of the defined oligonucleotide/ polynucleotides may be readily increased or decreased as desired or needed. . . . The length is generally guided by the principle that it should be of sufficient length to insure that it is on[] average only represented once in the population to be examined. [page 7]

Comparing the . . . hybridization patterns permits detection of those defined oligonucleotide/ polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions [of the solid support]. [page 13]

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained[;] rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used [to] form diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a 'master' solid support to which control sample and disease samples are alternately hybridized. [page 14]

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This[,] like many of the foregoing embodiments[,] may use known or unknown ESTs derived from many libraries. [page 18]

Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal[s], especially humans. [page 18]

WO 95/20681 ("Comparative Gene Transcript Analysis"), filed in 1994 by Appellants' assignee and published August 3, 1995, has three issued U.S. counterparts: U.S. Pat. Nos. 5,840,484, issued November 24, 1998; 6,114,114, issued September 5, 2000; and 6,303,297, issued October 16, 2001.

The specification describes the use of transcript expression *patterns*, or "images", each comprising multiple pixels of gene-specific information, for diagnosis, for cellular phenotyping, and in toxicology and drug development efforts. The specification describes a plurality of methods for obtaining the requisite expression data -- one of which is microarray hybridization -- and equates the uses of the expression data from these disparate platforms. In particular, and with emphasis added:

The invention provides a "method and system for quantifying the relative abundance of gene transcripts in a biological specimen. . . . [G]ene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens." [abstract]

"[W]e see each individual gene product as a 'pixel' of information, which relates to the expression of that, and only that, gene. We teach herein [] methods whereby the individual 'pixels' of gene expression information can be combined into a single gene transcript image,' in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood." [page 2]

"The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen. . . . The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts." [page 6]

"High resolution analysis of gene expression be used directly as a diagnostic profile. . . . " [page 7]

"The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed." [page 7]

"The invention . . . includes a method of comparing specimens containing gene transcripts." [page 7]

"The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens." [i.e., the results yield analogous data to microarrays] [page 8]

"Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made." [page 8]

"In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities." [page 9]

"In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a

method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens. . . . " [page 9]

"[T]wo or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells." [pages 9 - 10]

"The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens. . . . This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as 'gene transcript image analysis' or 'gene transcript frequency analysis'. The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism." [page 11]

"The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few." [page 12]

"[G]ene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates." [page 12]

"For example, gene transcript frequency analysis can be used to different normal cells or tissues from diseased cells or tissues. . . ." [page 12]

"In toxicology, . . . [g]ene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. . . . " [page 12]

"In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond." [page 12]

"In a further embodiment, comparative gene transcript frequency analysis is used . . . for the selection of better pharmacologic animal models." [page 14]

"In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a diseased state or condition." [page 14]

"An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined." [page 15]

"[T]his research tool provides a way to get new drugs to the public faster and more economically." [page 36]

"In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker." [page 38]

"[T]he gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript image analyses are evaluated as indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. . . . The . . . analysis highlights any toxicological changes in the treated patients." [page 39]

**U.S. Pat. No. 5,569,588** ("Methods for Drug Screening") ("the '588 patent"), issued October 29, 1996, with a priority date of August 1995, describes an expression profiling platform, the "genome reporter matrix", which is different from nucleic acid microarrays. Additionally describing use of nucleic acid microarrays, the '588 patent makes clear that the utility of comparing multidimensional expression datasets is independent of the methods by which such profiles are obtained. The '588 patent speaks clearly to the usefulness of such expression analyses in drug development and toxicology, particularly pointing out that a gene's failure to change in expression level is a useful result. Thus, with emphasis added,

The invention provides "[m]ethods and compositions for modeling the transcriptional responsiveness of an organism to a candidate drug. . . . [The final step of the method comprises] comparing reporter gene product signals for each cell before and after contacting the cell with the candidate drug to obtain a drug response profile which provides a model of the transcriptional responsiveness of said organism to the candidate drug." [abstract]

"The present invention exploits the recent advances in genome science to provide for the rapid screening of large numbers of compounds against a systemic target comprising substantially all targets in a pathway [or] organism." [col. 1]



"The ensemble of reporting cells comprises as comprehensive a collection of transcription regulatory genetic elements as is conveniently available for the targeted organism so as to most accurately model the systemic transcriptional response. Suitable ensembles generally comprise thousands of individually reporting elements; preferred ensembles are substantially comprehensive, i.e. provide a transcriptional response diversity comparable to that of the target organism. Generally, a substantially comprehensive ensemble requires transcription regulatory genetic elements from at least a majority of the organism's genes, and preferably includes those of all or nearly all of the genes. We term such a substantially comprehensive ensemble a genome reporter matrix." [col. 2]

"Drugs often have side effects that are in part due to the lack of target specificity. . . . [A] genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reports, the first compound is, a priori, more likely to have fewer side effects." [cols. 2 - 3]

"Furthermore, it is not necessary to know the identity of any of the responding genes." [col. 3]

"[A]ny new compound that induces the same response profile as [a] . . . dominant tubulin mutant would provide a candidate for a taxol-like pharmaceutical." [col. 4]

"The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters." [col. 4]

" A sufficient number of different recombinant cells are included to provide an ensemble of transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug. In a preferred embodiment, the matrix is substantially comprehensive for the selected regulatory elements, e.g. essentially all of the gene promoters of the targeted organism are included." [cols. 6-7]

"In a preferred embodiment, the basal response profiles are determined. . . . The resultant electrical output signals are stored in a computer memory as genome reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus or drug. This information is indexed against the matrix to form reference response profiles that are used to determine the response of each reporter to any milieu in which a stimulus may be provided. After establishing a basal response profile for the matrix, each cell is contacted with a candidate drug. The term drug is used loosely to refer to agents which can provoke a specific cellular response. . . . The drug induces a complex response pattern of repression, silence and induction across

the matrix . . . .The response profile reflects the cell's transcriptional adjustments to maintain homeostasis in the presence of the drug. . . . After contacting the cells with the candidate drug, the reporter gene product signals from each of said cells is again measured to determine a stimulated response profile. The basal o[r] background response profile is then compared with . . . the stimulated response profile to identify the cellular response profile to the candidate drug." [cols. 7-8]

"In another embodiment of the invention, a matrix [i.e., array] of hybridization probes corresponding to a predetermined population of genes of the selected organism is used to specifically detect changes in gene transcription which result from exposing the selected organism or cells thereof to a candidate drug. In this embodiment, one or more cells derived from the organism is exposed to the candidate drug in vivo or ex vivo under conditions wherein the drug effects a change in gene transcription in the cell to maintain homeostasis. Thereafter, the gene transcripts, primarily mRNA, of the cell or cells is isolated . . . [and] then contacted with an ordered matrix [array] of hybridization probes, each probe being specific for a different one of the transcripts, under conditions where each of the transcripts hybridizes with a corresponding one of the probes to form hybridization pairs. The ordered matrix of probes provides, in aggregate, complements for an ensemble of genes of the organism sufficient to model the transcriptional responsiveness of the organism to a drug. . . . The matrix-wide signal profile of the drug-stimulated cells is then compared with a matrix-wide signal profile of negative control cells to obtain a specific drug response profile." [col. 8]

"The invention also provides means for computer-based qualitative analysis of candidate drugs and unknown compounds. A wide variety of reference response profiles may be generated and used in such analyses." [col. 8]

"Response profiles for an unknown stimulus (e.g. new chemicals, unknown compounds or unknown mixtures) may be analyzed by comparing the new stimulus response profiles with response profiles to known chemical stimuli." [col. 9]

"The response profile of a new chemical stimulus may also be compared to a known genetic response profile for target gene(s)." [col. 9]

The August 11, 1997 press release from the '588 patent's assignee, Acacia Biosciences (now part of Merck) (reference 9h attached hereto), and the September 15, 1997 news report by Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," *Genetic Engineering News* (reference 9i attached hereto), attest the commercial value of the methods and technology described and claimed in the '588 patent.

WO 97/13877 ("Measurement of Gene Expression Profiles in Toxicity Determinations"), published April 17, 1997, describes an expression profiling technology differing

somewhat from the use of cDNA microarrays and differing from the genome reporter matrix of the '588 patent; but the use of the data is analogous. As per its title, the reference describes use of expression profiling in toxicity determinations. In particular, and with emphasis added:

"[T]he invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates." [Field of the invention]

"An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems." [page 3]

"Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals." [page 3]

"The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel . . . methodologies that permit the formation of gene expression profiles for selected tissues . . . . Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity." [page 3]

"As used herein, the terms 'gene expression profile,' and 'gene expression pattern' which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. . . . Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences determined in a gene expression profile is at least ten thousand." [page 7]

"The invention provides a method for determining the toxicity of a compound by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. . . . Gene expression profiles derived from test organisms are compared to gene expression profiles derived from control organisms. . . ." [page 7]

Therefore, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Evidence of the benefits of this information include:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers

disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.
- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

**C. Objective evidence corroborates the utilities of the claimed invention**

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Appellants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing millions of sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's

databases have proven to be valuable in, for example, the identification and development of drug candidates. Page et al., in discussing the identification and assignment of candidate drug targets, state that “rapid identification and assignment of candidate targets and markers represents a huge challenge ... [t]he process of annotation is similarly aided by the quantity and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals)” Page, M.J. et al., “Proteomics: a major new technology for the drug discovery process,” *Drug Discov. Today* 4:55-62 (1999) (Reference No. 5), see page 58, col.

2). As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte’s invention of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

Because the Patent Examiner failed to address or consider the “well-established” utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner’s rejections should be overturned regardless of their merit. Withdrawal of the rejection of claims 1-3 and 12-16 under 35 U.S.C. § 101 is therefore requested.

35 U.S.C. § 112, First Paragraph, Rejection of Claims 1-3 and 12-16

The Examiner has rejected claims 1-3 and 12-16 under 35 U.S.C. § 112, first paragraph, specifically, since the claimed invention is not supported by either a specific asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention.

Applicants Response

To the extent that this rejection under 35 U.S.C. § 112, first paragraph, is based on the improper allegation of lack of patentable utility under 35 U.S.C. § 101, for the reasons given by applicant above in response to that rejection, it fails for the same reasons and should therefore be withdrawn.

35 U.S.C. § 112, Second Paragraph, Rejection of Claims 1-3 and 12-16

The Examiner has further rejected claims 1-3 and 12-16 under 35 U.S.C. § 112, second

paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicants regards as the invention. These claims comprise subject matter drawn to non-elected inventions.

Applicants Response

Applicants submit that claim 1 properly recites a “combination” of polynucleotide sequences having the nucleic acid sequences of SEQ ID NOs:1-13, that includes the elected sequence of SEQ ID NO:6, and is therefore clear and definite. Claim 2 has been amended to recite only the elected sequence of SEQ ID NO:6 and claims 3 and 13-16 no longer depend from a claim reciting non-elected inventions. Claim 12 has been canceled. With these amendments and remarks, applicants submit that the claims are now clear and definite, and request withdrawal of the rejection of claims 1-3 and 12-16 under 35 U.S.C. § 112, second paragraph.

**CONCLUSION**

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding objections/rejections. Early notice to that effect is earnestly solicited. Applicants further request that, upon allowance of claims 1, 2 and 16, that claims 4-6, 10, and 18-20 be rejoined and examined as methods of use of the products of claims 1, 2 and 16 that depend from and are of the same scope as claims 1, 2, and 16 in accordance with *In re Ochiai* and the MPEP§ 821.04. Applicants also request that claim 17 be rejoined and examined as a composition of matter claim that depends from and further limits the composition of matter of claim 16.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact the undersigned at the number listed below.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108**.

Respectfully submitted,

INCYTE CORPORATION



David G. Streeter, Ph.D.

Reg. No. 43,168

Direct Dial Telephone: (650) 845-5741

Date: January 7, 2004

**Customer No.: 27904**

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555

Fax: (650) 849-8886

Attachment(s):

- 1) Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999);
- 2) Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).
- 3) John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, 107 Environ. Health Perspec. 107:681-685 (1999).
- 4) Email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.
- 5) Page, M.J. et al., Proteomics: a major new technology for the drug discovery process, Drug Discov. Today 4:55-62 (1999).
- 6) Declaration of John C. Rockett, Ph.D., under 37 C.F.R. § 1.132, with Exhibits A - Q;  
Declaration of Tod Bedilion, Ph.D., under 37 C.F.R. § 1.132;
- 8) Declaration of Vishwanath R. Iyer, Ph.D., under 37 C.F.R. § 1.132 with Exhibits A - E; and
- 9) ten (10) references published before the filing date of the instant application:
  - a) WO 95/21944, SmithKline Beecham, "Differentially expressed genes in healthy and diseased subjects" (Aug. 17, 1995)
  - b) WO 95/20681, Incyte Pharmaceuticals, "Comparative Gene Transcript Analysis" (Aug 3, 1995)
  - c) Schena et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science 270:467-470 (Oct 20, 1995)
  - d) WO 95/35505, Stanford University, "Method and apparatus for fabricating microarrays of biological samples" (Dec 28, 1995)
  - e) U.S. Pat. No. 5,569,588, Ashby et al., "Methods for Drug Screening" (Oct 29, 1996)
  - f) Heller al., "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays," PNAS 94:2150 - 2155 (Mar 1997)
  - g) WO 97/13877, Lynx Therapeutics, "Measurement of Gene Expression Profiles in Toxicity Determinations" (April 17, 1997)
  - h) Acacia Biosciences Press Release (August 11, 1997)
  - i) Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," Genetic Engineering News (Sept. 15, 1997)
  - j) DeRisi *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science 278:680 - 686 (Oct 24, 1997)



This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.

## IN PERSPECTIVE

Claudio J. Conti, Editor

# Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,<sup>1</sup> Michael Bittner,<sup>2</sup> Jeffrey Trent,<sup>2</sup> J. Carl Barrett,<sup>1</sup> and Cynthia A. Afshari<sup>1</sup>

<sup>1</sup>Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

<sup>2</sup>Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

**Key words:** toxicology; gene expression; animal bioassay

## INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

## MICROARRAY DEVELOPMENT AND APPLICATIONS

### cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

\*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22-24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25-27].

#### Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28-30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

#### THE USE OF MICROARRAYS IN TOXICOLOGY

##### Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

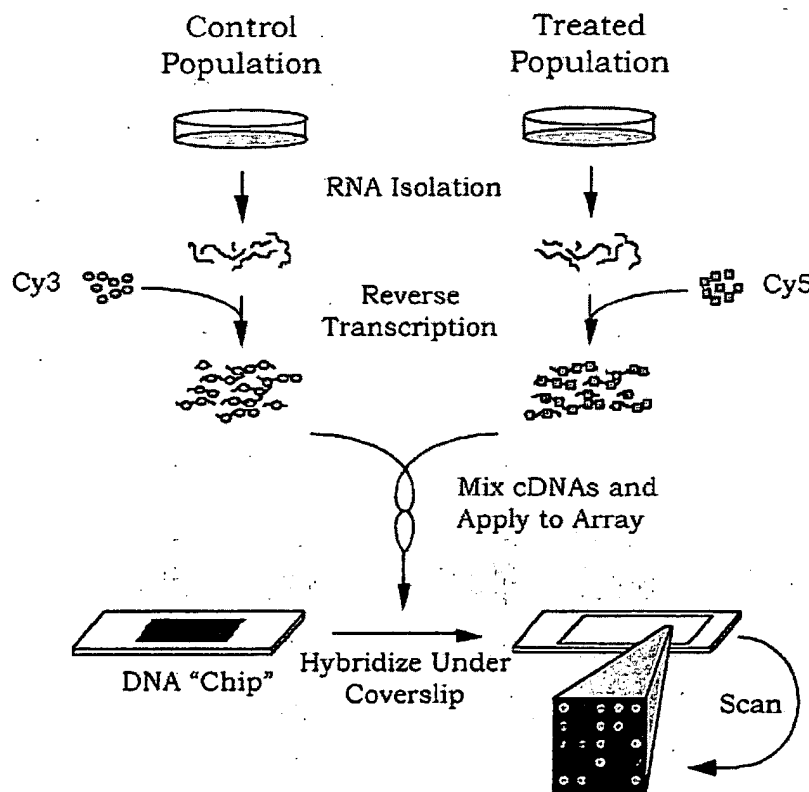


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

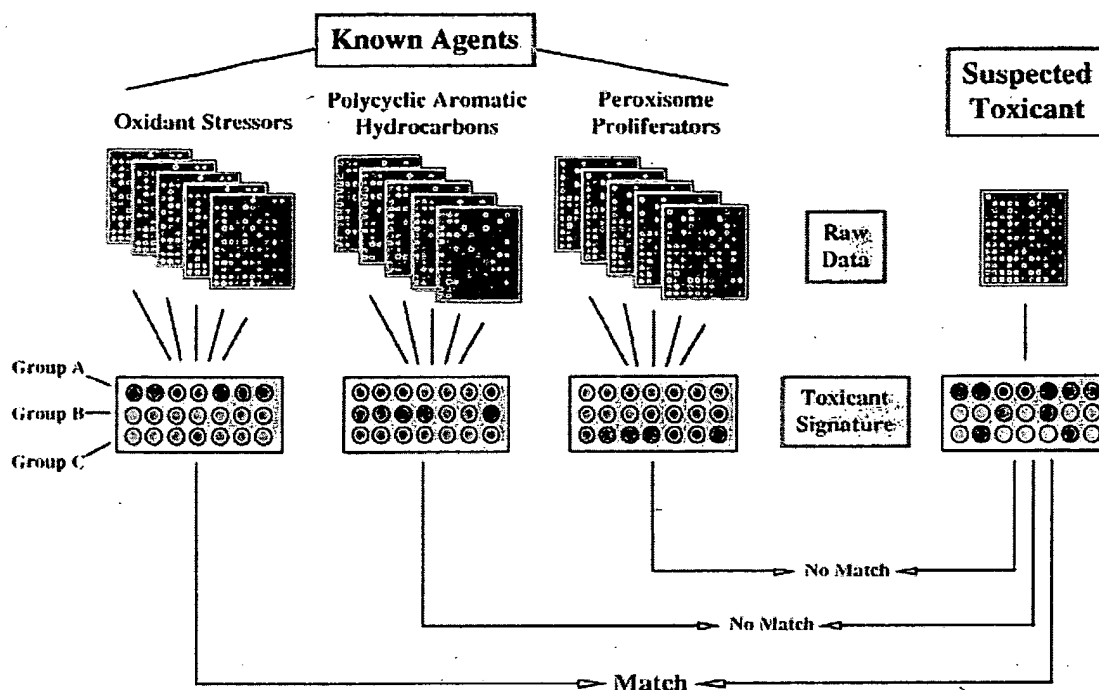


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

#### Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

\*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

#### Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

### Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

### FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

### ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

### REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>





## Expression profiling in toxicology — potentials and limitations

Sandra Steiner \*, N. Leigh Anderson

*Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA*

### Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

**Keywords:** Proteomics; Genomics; Toxicology

### 1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

\* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: [steiner@lsbc.com](mailto:steiner@lsbc.com) (S. Steiner)

## 2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

## 3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

## 4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

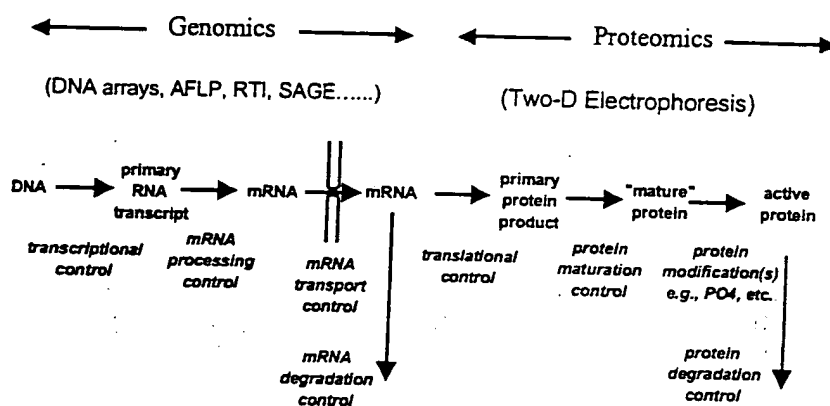


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

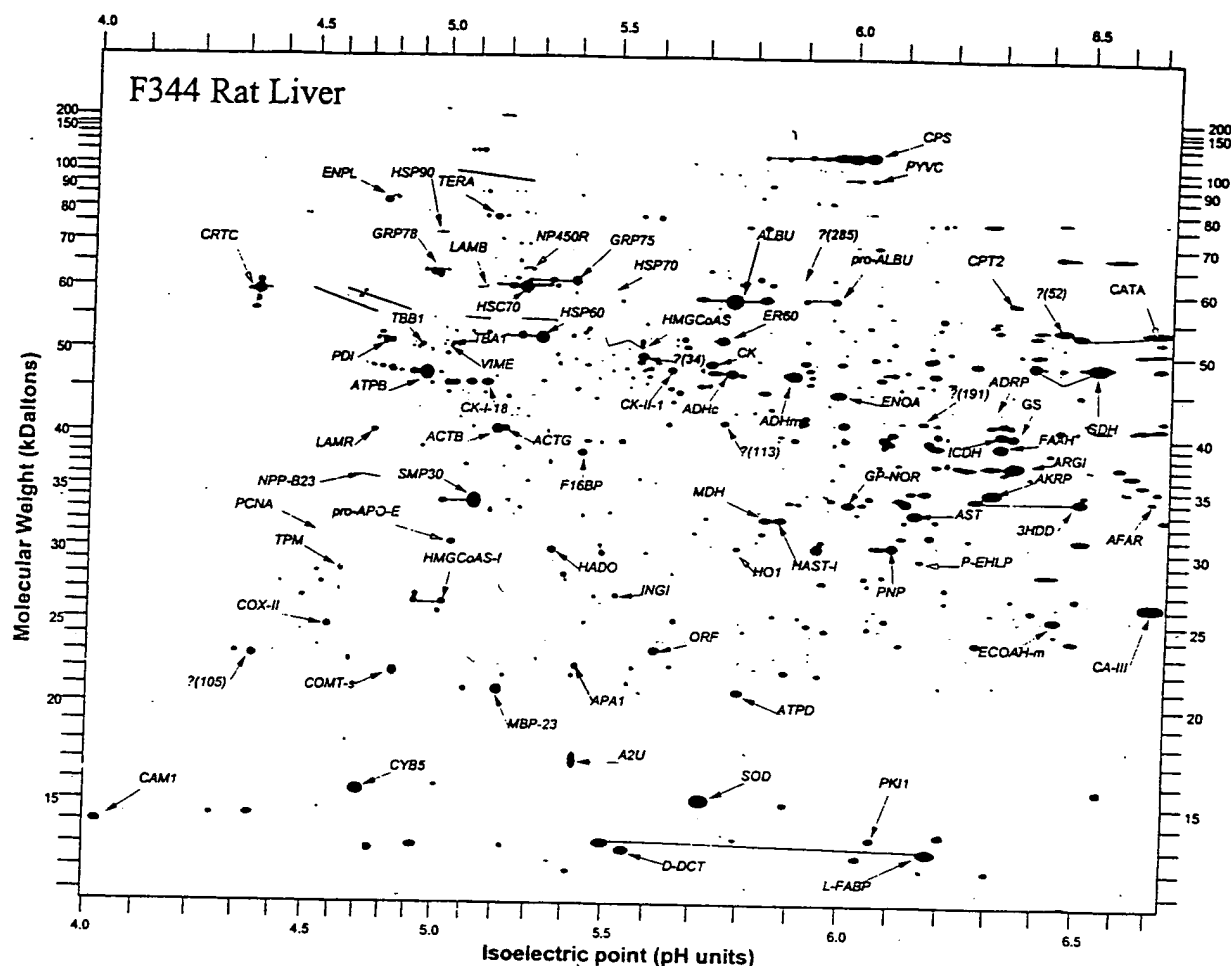


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

### 5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

## 6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

## 7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trials.

## References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610-614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355-363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767-773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338-345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157-161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467-470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639-645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raymackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *Biochem. Biophys. Res. Commun.* 218, 777-782.
- Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253-258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543-1544.

## Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

Docket No.: PB-0004-1 CIP

USSN: 09/818,143

Ref. No. 3 of 5

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7-8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. **Key words:** DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681-685 (1999). [Online 6 July 1999]

<http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html>

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (1). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer

has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7-8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

## Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in

a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (2). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

## Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: [rockett.john@epa.gov](mailto:rockett.john@epa.gov)

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic Microsystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrays, as described by Brown (3). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of  $> 2,500$  spots/cm<sup>2</sup> may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental

arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors affecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

### Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., <sup>32</sup>P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosphorimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptened, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (4). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA<sup>+</sup> RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After

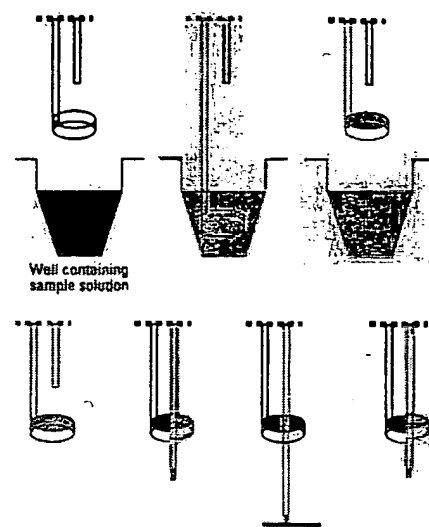


Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (14), with permission from Genetic Microsystems.

normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5' end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluorolabeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

## Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain  $> 10^8$  molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of  $< 1$  fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

## Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (5). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, C.

Table 1. Advantages and disadvantages of different microarray scanning systems.

Nonconfocal laser scanner			
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
Disadvantages	Fast scanning of bright samples		May have high light collection efficiency
	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
Resolution typically low			

CCD, charge-coupled device.  
From Kawasaki (13).

*elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes

transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

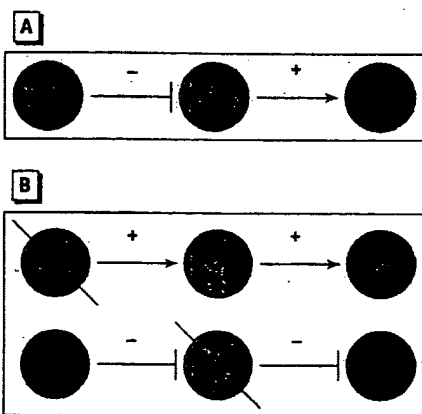


significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.



**Figure 2.** Potential effects of gene knockout within positively and negatively regulated gene expression networks.  $i_1$  is limiting in wild type for expression of  $i_2$ . (A) A simple, two-component, linear regulatory network operating on gene  $i_2$ , where  $i_1$  is a positive effector of  $i_2$  and  $j_n$  is either a positive or negative effector of  $i_1$ . This network could be deduced by examining the consequence of (B) deleting  $j_n$  on the expression of  $i_1$  and  $i_2$ , where the expression of  $i_2$  would be decreased or increased depending on whether  $j_n$  was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

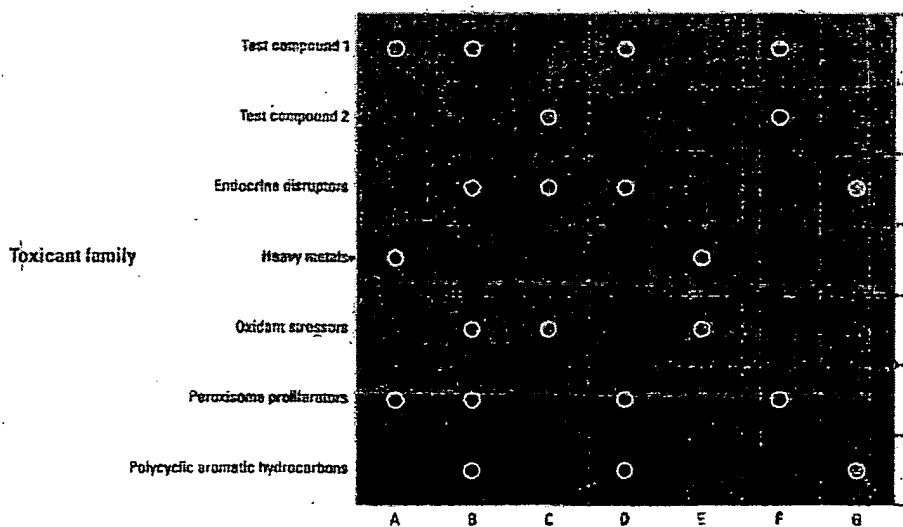
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, non-contaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a non-related species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings.

Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.

- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

## EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or



**Figure 3.** Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

## Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

## SPEAKERS

Cindy Afshari  
NIEHS  
Linda Birnbaum  
U.S. EPA  
Ron Butow  
University of Texas  
Southwestern Medical  
Center  
Alex Chenchik  
Clontech Laboratories, Inc.  
David Dix  
U.S. EPA

Abdel Elkhouloun  
Research Genetics, Inc.  
Sue Fenton  
U.S. EPA  
Norman Hecht  
University of Pennsylvania  
Pat Hurban  
Paradigm Genetics, Inc.  
Bob Kavlock  
U.S. EPA  
Emie Kawasaki  
General Scanning, Inc.

Steve Krawetz  
Wayne State University  
Nick Mace  
Genetic Microsystems, Inc.  
Scott Mordecai  
Affymetrix, Inc.  
Kevin Morgan  
Glaxo Wellcome, Inc.  
Elaine Poplin  
Research Genetics, Inc.  
Don Rose  
Cartesian Technologies, Inc.

Jim Samet  
U.S. EPA  
Sam Ward  
University of Arizona  
Jeff Welch  
U.S. EPA  
Reen Wu  
University of California  
at Davis  
Tim Zacharewski  
Michigan State University

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals; and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach

is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

## REFERENCES AND NOTES

1. The chipping forecast. *Nat Genet* 21(Suppl 1):3-60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: [www.ncbi.nlm.nih.gov/Schuler/UniGene](http://www.ncbi.nlm.nih.gov/Schuler/UniGene) [cited 22 March 1999].
3. Brown P.O. The Brown Lab. Available: <http://cmgm.Stanford.edu/pbrown> [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* 51:313-324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: [www.mcb.arizona.edu/wardlab/microarray.html](http://www.mcb.arizona.edu/wardlab/microarray.html) [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 4:1293-1301 (1998).
7. Brown P.O. The Full Yeast Genome on a Chip. Available: <http://cmgm.Stanford.edu/pbrown/yeastchip.html> [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog* 24(3):153-159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. *Bioessays* 20:555-561 (1998).
10. Zacharewski TR, Timothy R. Zacharewski. Available: [www.bch.msu.edu/faculty/zachar.htm](http://www.bch.msu.edu/faculty/zachar.htm) [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. *Mol Hum Reprod* 3:473-478 (1997).
12. Stipp D. Gene chip breakthrough. *Fortune*, March 31:56-73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: <http://www.geneticmicro.com/resources/html/coldspring.html> [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

Subject: RE: [Fwd: Toxicology Chip]

Date: Mon. 3 Jul 2000 08:09:45 -0400

From: "Afshari, Cynthia" <afshari@niehs.nih.gov>

To: "Diana Hamlet-Cox" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at  
<http://manuel.niehs.nih.gov/maps/quest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80-) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80- genes across our experiments. Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be. I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox  
> Sent: Monday, June 26, 2000 8:52 PM  
> To: afshari@niehs.nih.gov  
> Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.

> Can you help me in this matter? I don't need to know the sequences, necessarily, but I would like very much to know what types of sequences are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----

> Subject: Toxicology Chip  
> Date: Mon. 19 Jun 2000 18:31:48 -0700  
> From: Diana Hamlet-Cox <dianahc@incyte.com>  
> Organization: Incyte Pharmaceuticals  
> To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as pharmacotoxicology markers, and found the Press Release dated February 29, 2000 regarding the work of the NIEHS in this area. I would like to know if there is a resource I can access (or you could provide?) that would give me a list of the 12,000 genes that are on your Human ToxChip Microarray. In particular, I am interested in the criteria used to select sequences for the ToxChip, including any control sequences included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.  
> Incyte Genomics, Inc.

> --

>

> =====

> This email message is for the sole use of the intended recipient(s) and  
> may contain confidential and privileged information subject to  
> attorney-client privilege. Any unauthorized review, use, disclosure or  
> distribution is prohibited. If you are not the intended recipient,  
> please contact the sender by reply email and destroy all copies of the  
> original message.

> =====

>  
>  
>

# Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlf, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

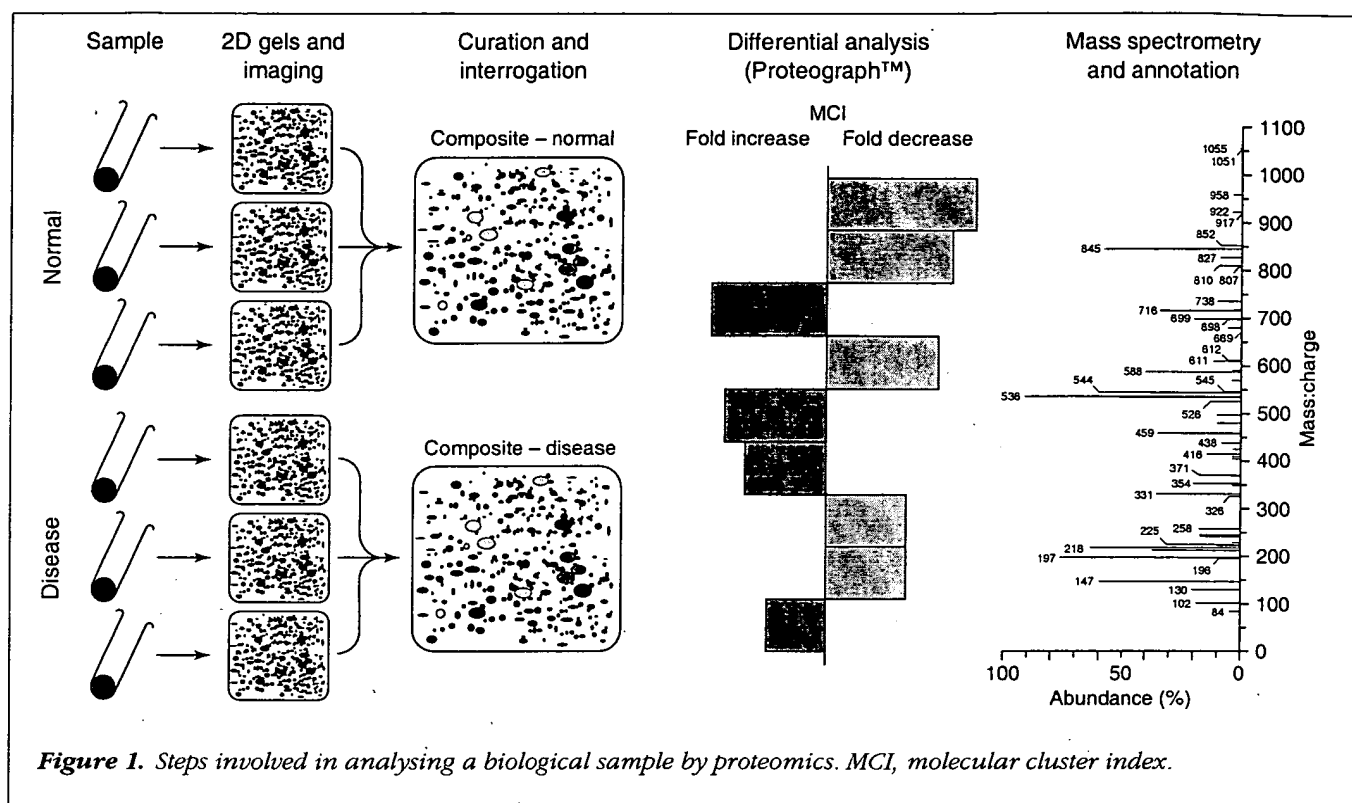
**A**mong the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should pro-

duce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic<sup>1,2</sup> and microarray<sup>3,4</sup> technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

**Martin J. Page\*, Bob Amess, Christian Rohlf, Colin Stubberfield and Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. \*tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk



analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

### Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed<sup>5-7</sup>. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins<sup>8</sup>, which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the

process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS-PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

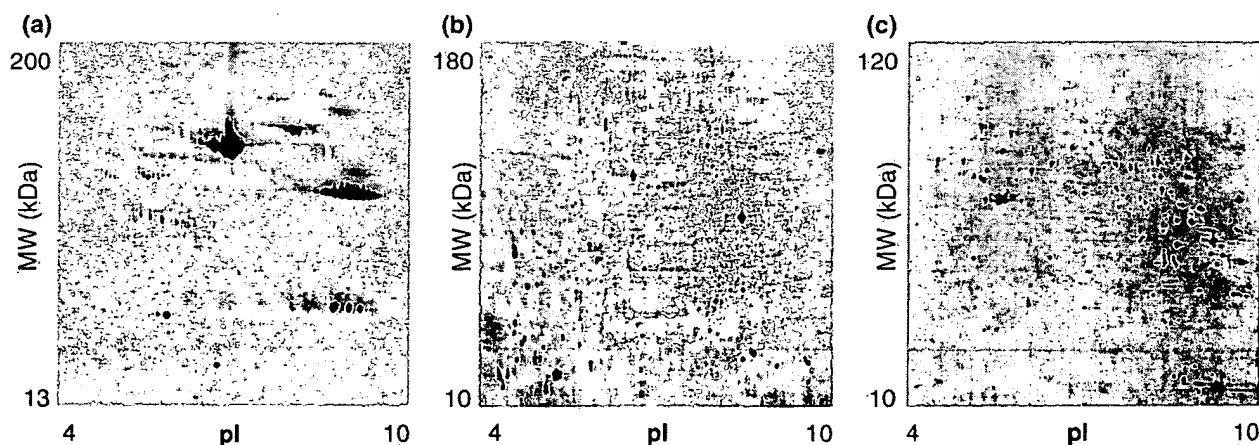
#### Use of proteomics to identify disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

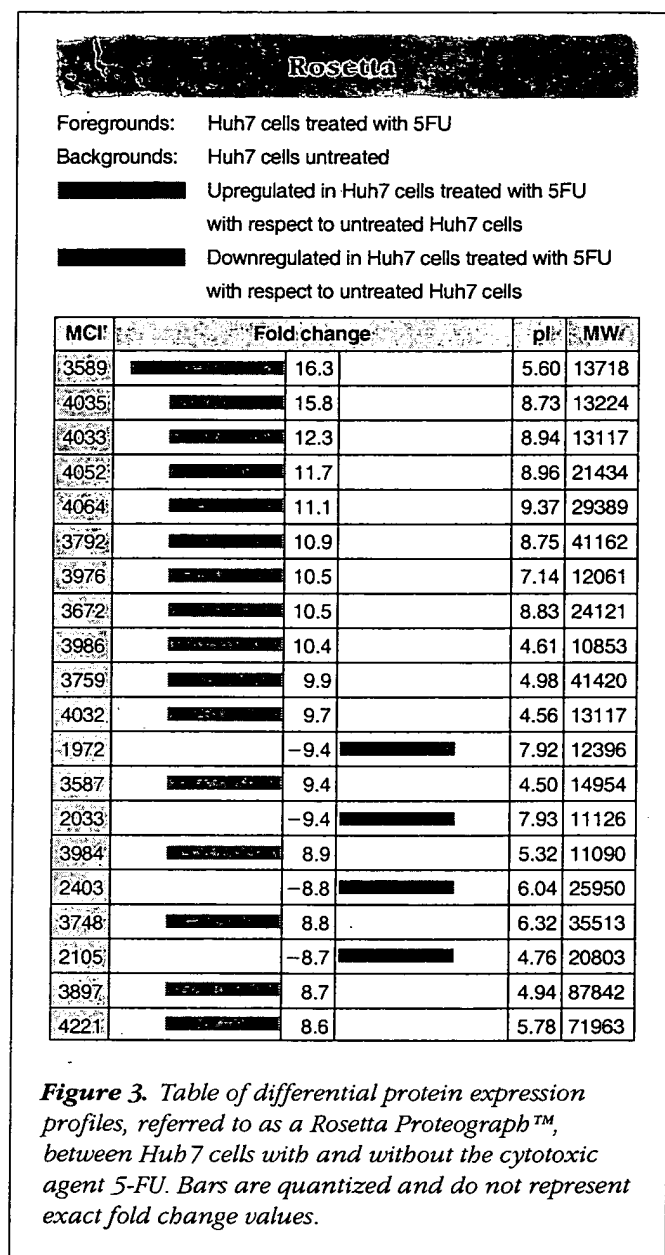
With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and im-

mediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical



**Figure 2.** Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus *Candida albicans* and (c) the human hepatoma cell line Huh7.



cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput

mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

#### Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry<sup>9</sup>. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas<sup>10</sup>, human breast proteins from normal and tumour sources<sup>11–13</sup>, lung tumours<sup>14</sup>, colon tumours<sup>15</sup> and bladder tumours<sup>16</sup>. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified<sup>17,18</sup>.

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.



### Proteomics for target validation and signal transduction studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences<sup>20</sup>.

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules<sup>21-23</sup>. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

#### Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics<sup>24,25</sup>. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affin-

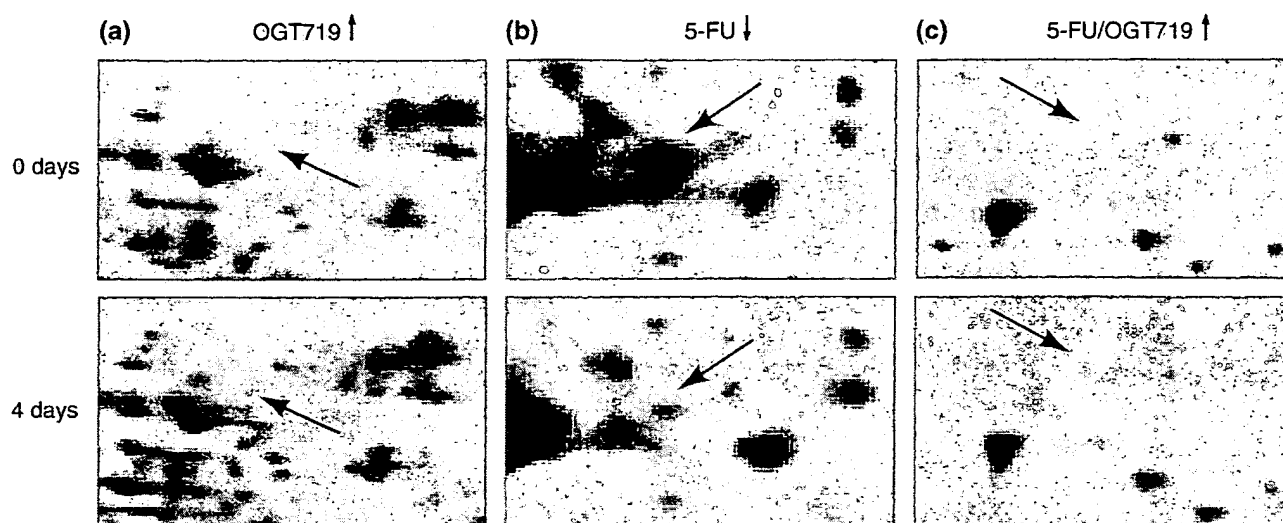
ity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies<sup>26-28</sup>. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1-10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

### Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable



**Figure 4.** Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1 $\alpha$ 2, (b) novel (three peptides by MS-MS) and (c)  $\alpha$ -subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.

of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes<sup>29</sup>, hepatoma Huh7 cells<sup>30</sup> and some colorectal tumour cells<sup>31</sup>. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with  $IC_{50}$  doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins.

Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein<sup>32</sup>, can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

### *Clear potential*

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

### **Use of proteomics in formal drug toxicology studies**

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members<sup>33,34</sup>, encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabo-

lism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

### *Unique P450 profiles*

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up<sup>35–37</sup>. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

### Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

### Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

### REFERENCES

- 1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
- 2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
- 3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
- 4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
- 5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
- 6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
- 7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
- 8 Parekh, R.B. and Rohlf, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
- 9 Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
- 10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
- 11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
- 12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
- 13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
- 14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
- 15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
- 16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
- 17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
- 18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
- 19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
- 20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
- 21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
- 22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
- 23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
- 24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
- 25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
- 26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
- 27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
- 28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
- 29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
- 30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
- 31 Mu, J.-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
- 32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
- 33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
- 34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
- 35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
- 36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
- 37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210